AFRL-OSR-VA-TR-2015-0176

A dedicated computational platform for Cellular Monte Carlo T-CAD software tools

Marco Saraniti
ARIZONA STATE UNIVERSITY

07/14/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTD
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
07-07-2015	Final	15-4-2014 to 14-4-2015

4. TITLE AND SUBTITLE
A Dedicated Computational Platform for Cellular Monte Carlo T-CAD Software Tools

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-14-1-0083

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Marco Saraniti

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Arizona State University
PO Box 876011
Tempe, AZ 85287-6011

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF, AFRL DUNS 143574726
Air Force Office of Scientific Research
875 N. Randolph St. Rm. 3112
Arlington, VA 22203-1954

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We report here on the acquisition of a specific computational platform with an optimized architecture for the Cellular Monte Carlo particle-based T-CAD simulation tools developed by our group. Such code is used for the modeling and design of electron devices realized with nontraditional semiconductor materials, where the experimental determination of material parameters for simulation purposes is arduous or incomplete.

The acquired equipment has been successfully deployed and is making possible the extraction of device parameters from simulation data performed at molecular resolution. For the first time, the parameters extracted with our tools will supply quantitative predictions of direct RF measurements as well as the electro-thermal device characteristics; the equipment will therefore allow the device design even in absence of reliable experiments.

15. SUBJECT TERMS
computational platform, Cellular Monte-Carlo, particle-based, T-CAD, simulation tool, electron devices

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Marco Saraniti
U	U	U	UU	24	19b. TELEPHONE NUMBER *(Include area code)* 480-965-2650

Reset

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

A dedicated computational platform for Cellular Monte Carlo T-CAD software tools

Marco Saraniti
Center for Computational Nanoscience
School of Electrical, Computer, and Energy Engineering
Arizona State University

Final Report for the AFOSR Grant FA9550-14-1-0083

1. Introduction

We offer here the final report about the acquisition of a non-standard computational platform specifically designed for, and dedicated to particle-based design and optimization of solid-state electron devices. The AFOSR Grant FA9550-14-1-0083 made the equipment acquisition possible.

The design of such specialized hardware is based on our decade-long experience in developing modeling techniques for and within defense-related projects managed by several agencies of the Department of Defense (DoD).

In particular, we identified a well-defined application space in which the particle-based paradigm plays a crucial role for the design, realization, and optimization of electron devices of strategic relevance for present and future military applications. To be fully developed (and explored) in a realistic amount of time, such application space requires a computational platform specifically designed for the algorithmic characteristics of particle-based simulation code. The target application space includes many scales both in time and space: it ranges from the nano-scale first principles used to develop the models implemented in the code, to the extraction of macroscopic quantities defining the performance metrics of high-frequency, high-power electron devices in a strictly unified framework. The acquired specialized equipment allows the extraction of device parameters from simulation data performed at molecular resolution, yet obtained in a realistic amount of time.

A complex hierarchy of modeling approaches has been defined within the last two decades in order to assist the design and optimization of electron devices. Indeed, most of the design work is currently performed by using standard commercial software tools [1,2] that allow the Computer Aided Design (CAD) of the whole manufacturing process from the wafer growth to the circuit (and package) level [3]. However, the adequacy of those tools is questionable when their use is attempted for the design and optimization of novel device structures realized with nonstandard semiconductor materials, for which a complete and robust parameter set is not available. Additionally, electro-thermal modeling of such devices is still in its infancy, and is currently based on rather draconian approximations. Last, the post-processing of particle-based simulation data has never been fully developed, especially for extreme high-frequency devices driven by large signals.

For such high-power, high-frequency applications, particle-based [4] simulation – rather than flux-based [5] – is most useful, because of the smaller set of parameters needed and their physics-based (as opposed to phenomenological) nature, which makes their determination a somehow less arbitrary and more robust process. The price to be paid for using the more accurate particle-based numerical models is two-folds: a *higher algorithmic complexity* during the implementation process and a *higher computational cost* when they

are employed for device design and optimization. Both these aspects have substantially limited the use of particle-based approaches for the realization of device structures in the industrial environment.

The main idea behind the particle-based simulation of electron devices is based on the fact that highly predictive and accurate simulations can be performed even when a detailed experimental characterization of the material properties -- or of the specific device family -- is not available. We have shown that, once the electronic structure and the phonon dispersion of the bulk material are available, an additional minimal set of parameters makes possible the reliable and highly predictive particle-based simulation of devices realized with any semiconductor in any 2D or 3D geometry that can be represented on a tensor-product grid. Differently from flux-based simulations, the particle-based approach supplies an exact solution, in statistical terms, of the semi-classical Boltzmann Transport Equation (BTE) [6]. We developed a simulative approach that reduces dramatically the time required for the extraction of the parameters used in the characterization of the material, and contextually allows for extremely fast simulations of the devices of interests. Our algorithmic approach is fully exploited by the systematic use of the acquired computational platform that, even if it is built with standard components, has substantial differences from the standard computer clusters currently used for device and process simulation.

Furthermore, the computational platform is a crucial component in the development of a novel approach for the self-consistent microscopic simulation of the electrical and thermal properties of semiconductor devices.

In summary, the acquired computational platform is being used for three main purposes:

1) The use of our existing software for the modeling of novel device structures and their characterization at the microscopic level.
2) The realization of novel particle-based simulation tools for the multi-scale modeling of electro-thermal properties.
3) The realization of data extraction and post-processing software tools coupled with our simulation software that complement and in some cases partially replace the rather difficult and expensive characterization techniques of devices operating at high power and high frequency.

The remaining part of this document is structured as follows: after an initial description of the particle-based full band approach that we use for device modeling and simulation, specific research projects are described in section 3, were we discuss recently completed projects funded by the DoD for the characterization of a new generation of Nitride-based field effect transistors. In particular, the effects of material properties were studied in relation to device reliability and longevity. A sub-project of the research effort described in section 3 is based on the large signal RF characterization of power devices, and has been funded by the Office of Naval Research (ONR) and it as been fully integrated in our computational framework.

The subsequent section 4 is devoted to the description of two research projects that are in their initial phase and are currently funded by the Defense Advanced Research Programs Agency (DARPA); their completion will be made possible by the acquired computational platform. Extension of the funding for these projects as described in section 5 has been proposed in 2014 and granted in year 2015.

Finally, a detailed discussion is provided of the enhanced research capabilities that the acquired computational equipment will make available to our group, as well as a technical description, including the expected longevity, of the equipment.

2. Particle-Based Simulation – The Full Band Cellular Monte Carlo Approach

The self-consistent Ensemble Monte Carlo (EMC) [7] algorithm has been employed now for over thirty years to simulate semi-classical charge transport in semiconductor materials and devices (see figure 1). The limitations of the approach have been the large computational burden of the technique, particularly when the full representation of the electron dispersion and the phonon spectra are incorporated into the physical description of the material system of interest. This computational burden has limited the EMC method to semiconductor device analysis rather than design applications, as well as calibration of less computationally demanding approaches based on the moment equations. Furthermore, because of the computational burden associated with particle-based approaches, the procedures for the extraction of parameters relevant for the device engineers from the microscopic simulations have been developed sporadically or not developed at all.

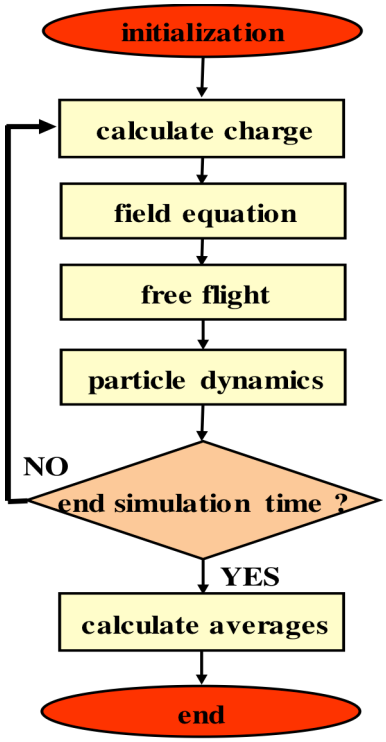

Figure 1. Flow-chart of particle-based simulation algorithms.

In order to reduce the computational demands of particle-based simulation, the cellular automaton (CA) approach was developed in the context of semiconductor device simulation [8]. In the CA approach, both **k**-space and position space are discretized, which simplifies the description of scattering and the particle motion in the phase-space. This technique was successfully demonstrated using an analytical, non-parabolic band model, where significant speed-up was observed compared to more traditional EMC methods.

This early work on CA methods utilized simplified band models to represent the energy dispersion and scattering rates, whereas state-of-the-art particle simulation techniques have increasingly moved towards full-band models. For this reason, we have developed a full-band CA-based simulator, which we briefly describe here. This simulation tool is based on the discretization of the first Brillouin Zone [9] (BZ) of the semiconductor crystal onto a non-uniform mesh, over which the transition table for the scattering probability for every initial state to every final state in this mesh is generated and stored. This leads to considerable simplification in the final state selection after scattering, which typically is a time-consuming process in full-band EMC simulation. Because the full **k**-space is utilized, completely anisotropic scattering rates may be incorporated without loss of performance, although at the cost of large memory usage. The large memory requirements result in a trade-off between accuracy of the final state energy and computer memory, which we address through the implementation of a non-uniform grid scheme.

The basic idea of our full-band CA approach [10] is to provide as accurate a physical description as possible compared to full-band EMC [11], while improving the computational efficiency of the method. This new approach is the algorithmic core of the Cellular Monte Carlo (CMC) [12] method.

CMC scattering algorithm. Within the full-band CMC framework, the scattering table is computed directly from perturbation theory [13], and the total probability of scattering due to all mechanisms is stored for each pair of **k**-space cells. In this way, the selection of the final state is reduced to the generation of a single random number to select the final state. The drawback of this approach is that a rather large scattering table is required to store all possible initial and final states in the first BZ. Furthermore, by storing only the total rate, information is lost about the exact type of scattering process involved in the transition (e.g. the nature of the phonon involved, and its energy).

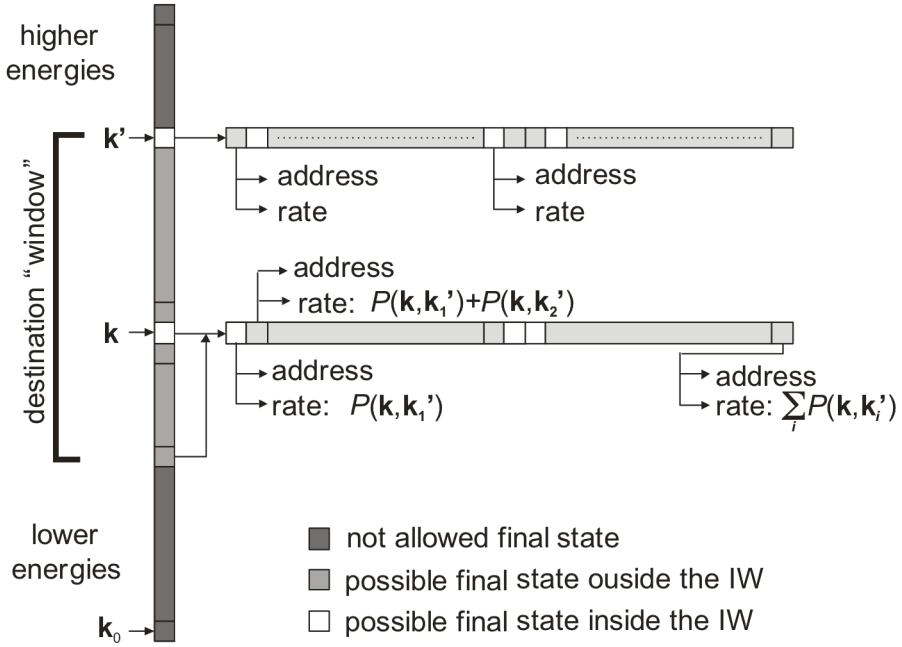

Figure 2. Schematic layout of the scattering table as stored in memory. Each **k**-point in the IW points to a set of possible destinations stored with their associated transition probability. The **k**-points outside the IW point to their equivalent state within the IW.

This lack of knowledge of the exact mechanism for scattering makes it impossible to refine the position of the carrier within the final momentum cell, which necessitates finer grids, or smaller cells, than the ones used by the aforementioned EMC algorithms, further exacerbating the memory requirements. In principle, separate scattering tables could be generated for each mechanism, at the cost of multiplying the total memory requirements by the number of mechanisms. In fact, this is done for certain mechanisms such as impact ionization. Through introduction of a non-uniform mesh in **k**-space, and through elimination of low-probability scattering events, and by implementing real-time compression techniques, the transition table may be reduced to manageable levels in terms of memory requirements compatible with present computational platforms.

Structure of the CMC transition table. The dimension of the CMC scattering table is much larger than that required by standard EMC methods, due to the storage of transition probabilities between all initial and final states in the BZ. There is an obvious trade-off between the accuracy of the final state energy (determined by energy conservation

requirements), and the amount of active memory (RAM) which is dependent on the size of the **k**-space mesh. Since the dimension of the table is proportional to the square of the number of cells used to represent BZ, a coarse grid would produce a relatively small table but would violate energy conservation, while an extremely fine grid would generate a transition table too large to be stored in the memory of the computer [14].

The physical nature of the system is helpful in addressing this memory problem. First, one can observe that the amount of error in energy conservation that can be considered "acceptable" strongly depends on the position within BZ. In Si for example, a very fine grid is required around the highly populated, low energy minima of the first conduction band, as well as around the Gamma point in the three upper valence bands. On the other hand, regions of the BZ with energy above 2.5 eV will be scarcely populated, as impact ionization is very efficient in depopulating high energy carriers. The implementation of an irregular grid in **k**-space is therefore crucial in optimizing the memory usage and making the approach possible.

Furthermore, symmetry within the BZ can be exploited by tabulating the transition probability only for the cells inside the Irreducible Wedge (IW) of the BZ, so reducing, for example, the final dimension of the transition table by a factor of 48 for Diamond crystal structures. Figure 2 represents the structure of the CMC transition table. For each band, cells representing the initial states within a given energy range are tabulated. In those cells, located inside the IW, a pointer is set to all possible final states, identified by its address (*i.e.* its location in BZ) and the associated rate. Cells outside the IW use the final state array of the corresponding cell in the IW. A rather small price has to be paid due to the need to rotate any initial **k** vector to the IW (if it is outside) prior of scattering, and back-rotate the computed final state **k'** via the inverse rotation. The first operation is quickly performed since a pointer is stored for each cell corresponding to the cell in the IW. The back-rotation is a little more complex, but equally fast. The inverse of the 3×3 rotation matrix is also stored for each grid-point, and used at the end of the scattering selection process to project the vector **k'** to the proper position.

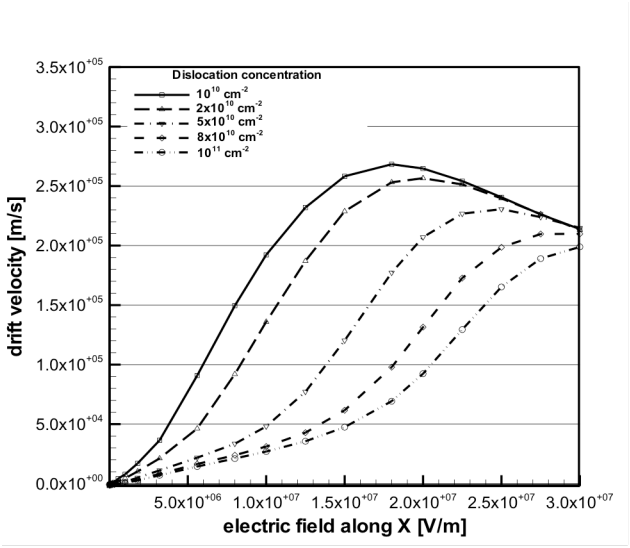

Figure 3. Velocity-Field curves for different dislocation concentrations, the free carrier concentration is set to 10^{17} cm^{-3}.

Multigrid Solver for Elliptical PDEs. Besides scattering, the second algorithmic bottleneck of particle-based simulation is the Poisson solver (see figure 1), which is executed over the whole simulation domain at least every femtosecond of simulated time.

Because of the need for repeated solutions of Poisson's equation and the obvious availability of the previous potential distribution to be used as a guess for the current solution, we implemented the iterative version of the multigrid algorithm [15] in the 2D [16] and 3D [17] Poisson solvers included in our code. Since the approach is based on a

hierarchy of grids that allows the simultaneous suppression of all the Fourier components of the error during the relaxation process [18], the performance of the multigrid algorithm scales linearly with the number of grid points. We demonstrated the robustness of our approach in the presence of irregularly shaped equipotential surfaces, dielectric interfaces, large gradients of charge, and in-homogeneous albeit tensor-product grids. The performance of our multigrid Poisson solver is more than satisfactorily; furthermore, to the best of our knowledge, there aren't many alternatives to the approach we have chosen if one wants to keep the geometrical generality of the computational domain, and a tensor-product grid with non-periodic boundary conditions. Indeed, a few alternatives approaches may perform slightly better than multigrid (see, for example, the FACR method [19]), but carry a cost in terms of restrictions on the geometry that we cannot afford, if we want to simulate the systems of interest with an high degree of geometric realism. The little popularity (relatively speaking) of the multigrid approach may be explained perhaps by the difficulty of its implementation [20], which usually results in extremely long and difficult to debug Poisson solvers, making the multigrid method hardly suited for academic software development. As a final remark about performance, we want to address the issue of parallel performance of multigrid algorithms. We did not write a parallel version of our multigird Poisson solver. Indeed, a general implementation of the algorithm is definitely arduous since it depends on the complexity of the error relaxation scheme adopted by the algorithm. In particular, while checkerboard relaxation is naturally suited for parallel implementation (as opposed as line relaxation or ILU relaxation [21][22]), it is well known that its performance is poor in the presence of in-homogeneous grids with non-cubic grid cells [23] (this is another reason that reduces the popularity of multigrid algorithms: handling irregular grids while conserving performance is more complicated from a programming viewpoint).

3. Recent DoD Funded Research on III-N Materials, Devices and Circuits.

We describe here the results of a research project funded by the Air Force Research Laboratory (contract number FA8650-08-C-1395, monitor: Dr. Christopher Bozada, and subsequent contracts). The project aimed to the modeling and simulation of GaN High Electron Mobility Transistors (HEMT) [24], with a particular emphasis on the effects of material defects on the RF performance and longevity of the devices. This research effort was targeted to the realization and validation of robust and predictive models of those material defects that affect the long-time reliability of devices. In strict collaboration with an experimental group at the Physics Department of ASU (PIs: Dr. Martha McCartney and Dr. David Smith), the PI Saraniti has been working at the realization of quantitative numerical models of the effects of material defects on device performance. While this research project has been recently completed, the larger memory size of the acquired equipment makes possible the extension of the RF

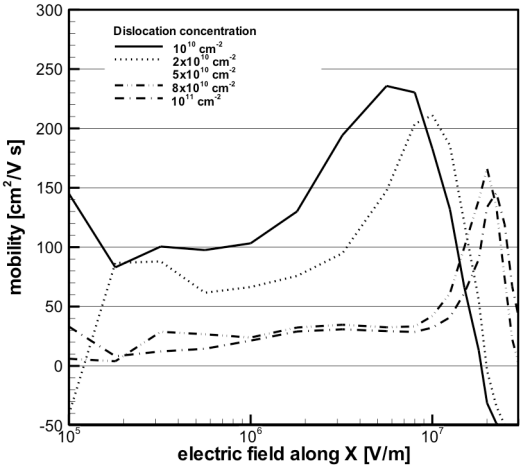

Figure 4. Effect of dislocation scattering on the mobility of GaN with doping concentration of 10^{17} cm^{-3}.

analysis to the highly nonlinear large signal regime, and ensures a significant increase of the accuracy of the results.

Bulk GaN. Devices based on AlGaN/GaN heterojunctions show excellent performance in terms of high drain current, low noise, and acceptably high cut-off frequencies. Because of these properties, GaN HEMTs are suitable for high power and high frequency applications. In particular, GaN HEMTs are currently in production for high power applications in satellite communications technology and local multi point distribution systems in the frequency range of 25-40 GHz and higher.

We performed a complete analysis of the transport properties of the bulk GaN material through our CMC simulator in order to identify and characterize the parameters affecting the device performance.

Usually, AlGaN/GaN epilayers are grown on sapphire or SiC substrates, and the resulting high lattice mismatch (between 2.5% and 12%) for hexagonal systems such as wurtzite GaN leads to a high level of dislocation density in the 10^7-10^{11} cm^{-2} range [25]. Dislocations are oriented along the *c* axis of the material and act as traps that can capture electronic charges and become Coulomb scattering centres. In this study, we assumed that the fraction of filled traps varies in the [0.5,0.6] range, accordingly to the free carrier concentration. This value is appropriate for a dislocation density about $5x10^{10}$ cm^{-2} in the range of doping concentration between 10^{15} and $5x10^{17}$ cm^{-3}. The aim of this work was to model how the presence of dislocations affects the device performance and to predict the limit value of dislocation density that allows normal device operation.

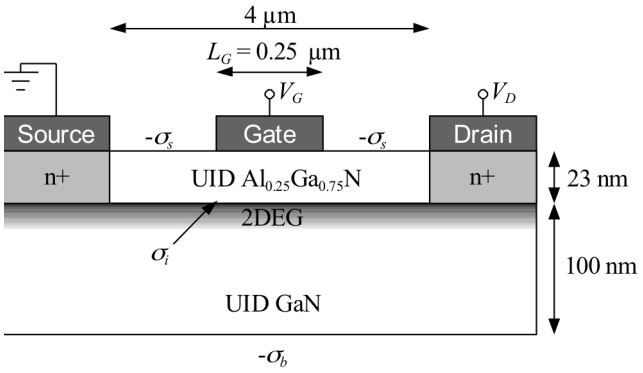

Figure 5. Simulated 250 nm AlGaN-GaN HEMT. The surface charge σ_S is used to reproduce the surface potential, the sheet charge, σ_b is chosen to fit the measured threshold voltage, while σ_i is used to model the piezoelectric effect.

In order to evaluate the effects of the dislocations with full band Monte Carlo simulation, the dislocation scattering rate ($1/\tau_{dis}$) has been expressed using the approach of Weimann *et al.* [26]:

$$\tau_{dis} = \frac{\hbar^3 \varepsilon^2 c^2}{N_{dis} m^* e^4} \frac{\sqrt{(1+\lambda^2 k_\perp^2)}}{\lambda^4} \qquad (1)$$

where N_{dis} is the concentration of threading edge dislocations, $\hbar$ is the reduced Plank's constant, m* is the effective mass, e is the electron charge, ε is the dielectric constant, $k_\perp$ is the incoming wave vector, and λ is the screening length.

In figure 3 we can see the influence of dislocations on the carrier velocity in bulk GaN. As expected, the scattering rate increases as the dislocation concentration increases.

The reduction of mobility due to dislocation scattering results in a smaller velocity response to the electric field. Furthermore, the carrier velocity does not show a linear behaviour at low fields, especially when the concentration of dislocations is high.

The low field differential mobility is shown in figure 4, where the simulations were performed by fixing the doping concentration at 10^{17} cm^{-3} and by using the dislocation concentration as parameter. For increasing electric field, the differential mobility increases until it reaches a peak (around 10^7 V/m), and subsequently decreases with a pronounced negative slope. Transport is dominated by phonon scattering for the region after the maximum mobility peak, while the scattering with charged dislocations influences the transport at low fields. In this regime, the probability of interaction between the carriers and the charged dislocations is reduced as the particle velocity increases due to the coulombic nature of the interactions, resulting in an increasing differential mobility with the field.

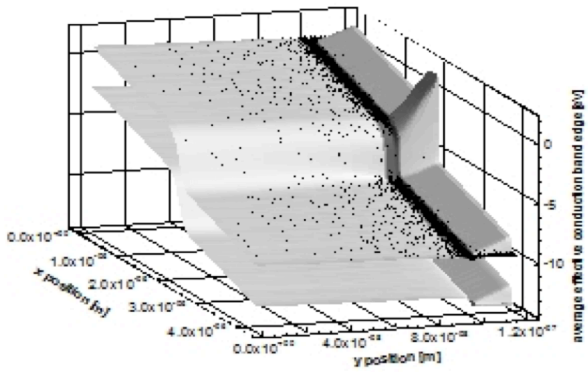

Figure 6. Band diagram and electron distribution in the simulated HEMT, for a bias of V_G=2V and V_{DS}=10V.

GaN Devices. AlGaN-GaN HEMTs have demonstrated a number of advantages in high power applications and high temperature operations. Compared to GaAs and InP HEMTs, GaN HEMTs show higher break-down voltage and thermal properties. The simulated layout (see figure 5) corresponds to the device described by Lee *et al*. in [27] and consists of a 100 nm layer of unintentionally doped GaN and 23 nm of $Al_{0.25}Ga_{0.75}N$. The unintentional doping value of 10^{17} cm^{-3} was assumed for both the GaN substrate and the AlGaN layer. The regions under the drain and source contacts were n-doped 10^{19} cm^{-3}. A surface charge donor layer σ_i, is located between the AlGaN and GaN layers in order to reproduce the piezoelectric effect at the heterojunction. In accordance with [28], we expressed the total polarization of the GaN and AlGaN layers as the sum of the spontaneous polarization P_{SP} and the strain induced by the heterojunction or piezoelectric polarization P_{PE}.

The sheet charge density reproducing the effects of polarization at the heterojunction has therefore been computed through the following expression:

$$\sigma(x) = P_{PE}(Al_xGa_{1-x}N) + P_{SP}(Al_xGa_{1-x}N) - P_{SP}(GaN)$$

that gives the value of -0.022 C/m^2 used for the simulations.

As depicted in figure 5, the gate length was 250 nm and the distance between source and drain was 4 μm. On the bottom of the device a sheet charge σ_b was set to -0.001 C/m^2 in order to fit the measured threshold voltage, while a surface charge of σ_s=-0.008 C/m^2 was used to reproduce the surface potential of the AlGaN layer which was 2 eV [29].

In Figure 6, the conduction and valence band profile of the simulated HEMT are shown with the charge distribution corresponding to a bias of 2V on the gate and 10V between drain and source. The black points represent a snapshot of the simulated electron population at steady-state. The current-voltage characteristic of the device was successfully validated with the experimental data.

In Figure 7 the I_d-V_g is shown. The extrapolated threshold voltage was -3.75 V in excellent agreement with the experiment.

In order to compute the screening length used for the dislocation scattering rate (equation 1), the net carrier concentration in the channel is needed. For HEMT devices, the transport occurs mostly in proximity of the heterojunction, hence we computed the free carrier concentration within the channel in order to accurately reproduce the effects of dislocations along the conductive path. We then performed an average of the carrier number over different bias values resulting in an effective free carrier concentration of 1.2×10^{19} cm^{-3}. We performed a series of simulations by changing the dislocation concentration within the device while keeping all the other parameters constant. By increasing the dislocation concentration, the device was driving less current due to the strongly reduced electron mobility.

The device performance was considerably compromised at dislocation concentration of 5×10^{11} cm^{-2}. At this concentration, the fraction of filled traps was high, and almost all the dislocations were activated and they acted as scattering centers.

The effect of dislocations was less detrimental when the concentration was about 10^{9} cm^{-2}. Indeed, with decreasing dislocation concentration, the output current increases but the I_d-V_d curves still saturate when the dislocation concentration is less than 10^{10} cm^{-2}. In agreement with the statistical occupation of traps computed by Weimann, the number of filled traps for a dislocation concentration of 10^{10} and 10^{9} cm^{-2} is the same for a free carrier concentration equal to 10^{19} cm^{-3}. Hence, the difference for the I_d-V_d and I_d-V_g curves at dislocation concentration of 10^{9} and 10^{10} cm^{-2} is negligible.

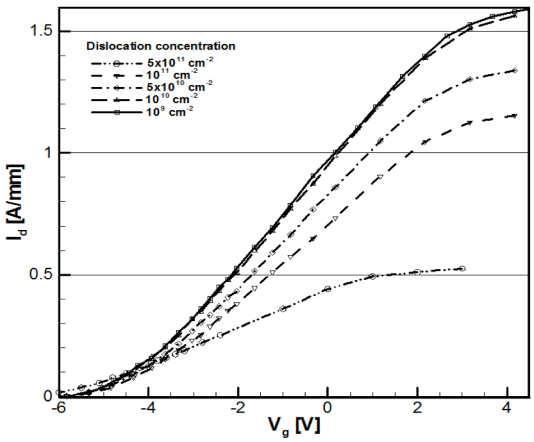

Figure 7. Id-Vg characterization for different values of dislocation concentration, Vd=10V.

Large-Signal AC Simulation.

Under the funding mentioned above, a detailed feasibility study has been performed for a project aiming to the integration of the CMC simulation software with various post-processing tools allowing the extraction of time-domain and frequency-domain parameters to be used in the optimization stage of the device design as well as in circuit simulations.

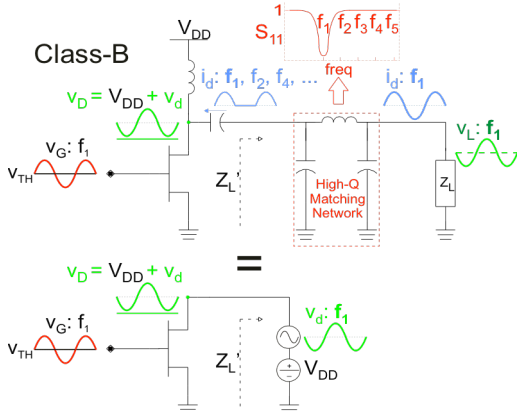

Figure 8: Example of a Class-B amplifier with high-Q matching network emulated by an active load-line technique.

The results of such initial study are reported here, while a proposal has been recently funded by the ONR to fully integrate in our code a new functionality for the large signal RF characterization of electron devices. It should be noted that the extraction of RF parameters is paramount for the understanding of

the large-signal, high-frequency operations of the devices and, to the best of our knowledge, have never been attempted in the framework of particle-based full-band simulations. The computational platform is instrumental in coupling the high algorithmic efficiency of the CMC algorithm with the complex procedures for data extraction.

Traditional small-signal AC analysis fails to predict large-signal device performance, which must be assessed by simulating the device within the full range of actual operating conditions. Thus, the high-Q matching network (i.e. highly selective in frequency) of a power amplifier must be included within the simulation framework in order to properly simulate the dynamic load-line (i.e. the drain voltage swing due to the load) seen at the device output. A time domain solution of these matching networks including large reactive elements necessarily implies long transient times that make particle-based simulation unpractical. However, this issue can be overcome by connecting a sinusoidal voltage generator, tuned at the fundamental frequency, at the device output. This emulates the load-line drain voltage swing at the fundamental harmonic, and presents a short circuit at the other harmonics, effectively emulating a high-Q matching network as shown in figure 8.

The actual synthesized load impedance is determined in post-processing through a Fourier transform. The magnitude and phase of the complex load can be also adjusted, by changing magnitude and phase of the voltage generator, and the simulation and the subsequent impedance analysis can be iterated until the desired load impedance is obtained. In such way, we can emulate a constant load for different input powers, and characterize a simulated device under large-signal operations to obtain typical figures of merit as shown in figure 9. Moreover, the analysis of the dynamic load-line, shown in figure 10, can provide valuable insight regarding the device large-signal operations [30] as directly related to the carrier dynamics.

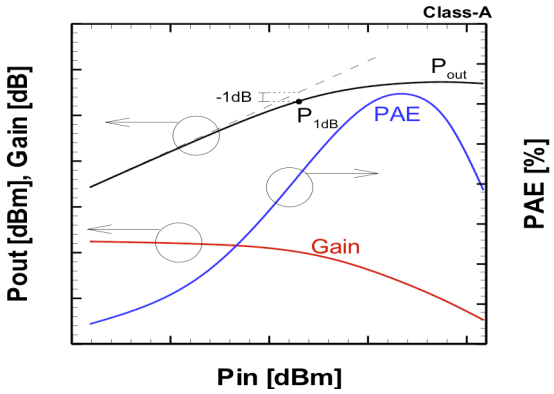

Figure 9. Typical large-signal figures of merit of a FET Class-A power amplifier

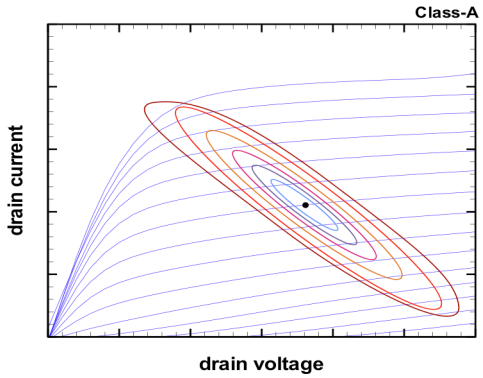

Figure 10. Typical dynamic load-lines of a FET Class-A power amplifier for increasing input power

In such a way, the performance under large-signal operation can be evaluated through a physical device simulator, when the device stimuli appropriately mimic the actual device operating bench. All the large-signal non-stationary non-linear effects are intrinsically included in the microscopic models of particle-based carrier dynamics modeled by the CMC algorithm, which offers a more accurate description than the conventional commercial drift and diffusion simulators.

The Harmonic Balance Algorithm. The iterative procedure described in the previous paragraph is a particular single-frequency case of the general frequency-domain circuit

solver known as Harmonic Balance (HB). An automated version of the described procedure can be implemented, self-consistently coupled with our CMC device simulator, and extended to a variable number of harmonics allowing the inclusion of any kind of impedance and network connected to any of the device contacts as shown in the equivalence of the two schematics depicted in figure 11.

Unlike previous Monte Carlo codes coupled with time-domain circuit solvers used for FET power analysis [31], our CMC/HB simulator allows a time-efficient simulation of devices connected to high-Q matching networks (required to suppress undesired harmonics generated due to non-linearity) typical of mm-wave band power amplifier classes such as -AB, -B, -F, and hard-driven Class-A. Due to the iterative nature of the HB algorithm, this simulation approach is possible because of the generally higher computational efficiency of our CMC code as compared to the traditional EMC approach. On the other hand, a frequency domain circuit solver such as the HB allows the simulation of the steady-state behavior of a transistor and its external passive reactive network without the need for simulating a long transient time (due to large *RC* and/or *L/C* time constants) typical of time domain solutions.

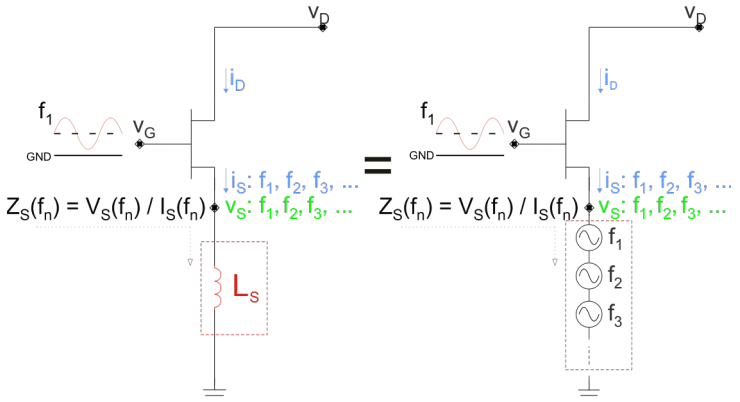

Figure 11. Example of the HB algorithm: the source inductor is emulated by generating the voltage sinusoids

Large-Signal Characterization of mm-wave GaN Power Amplifier. By exploiting the self-consistently coupled CMC/HB device simulator, we can predict the large-signal performance of mm-wave FET power amplifiers and efficiently characterize the RF power performance of state-of-the-art high-power, high-frequency devices like GaN HEMTs. In such way, we can relate reliability issues and material defects to the large-signal power performance of mm-wave GaN HEMTs by simply including new/improved physical models in our CMC device simulator. In this way, large-signal characterization can be readily and conveniently available during the device layout design phase. The goal of these large-signal Monte Carlo device simulations is to obviate the inability of properly evaluating analytically the large-signal performance, providing a TCAD tool for the device early-stage design flow. This is done by providing a computer-based performance evaluation in lieu of the extremely time-consuming and expensive iterations of prototyping and experimental large-signal characterization. This simulated large-signal performance evaluation will include the output signal spectrum in non-linear regime, and typical large-signal figures of merit, such as 1dB compression gain (P1dB) and third-order interception point (IIP3).

The typical characterization flow, in the case that preliminary experimental DC and small-signal RF characterization is available for the family of devices under investigation, initially starts with a fit of the experimental DC characteristics. Then the RF small-signal performance are also assessed through CMC/AC simulations (by applying small step perturbations in order to extract a two-port network characterization of the device through Y-parameters), and the agreement with the experimental measurements can be verified. At this point, once the agreement between simulated and experimental DC/RF small-signal performance is assessed as starting baseline, we finally proceed with the Power Amplifier

characterization through the CMC/HB code. In particular, our investigation will focus on the optimization of the intrinsic device layout as well as real device issues such as parasitic elements (i.e. gate resistance, source/drain inductors), material defects/reliability (i.e. threading dislocations, surface traps, interface roughness), and self-heating/thermal management.

4. Current Research on Electro-Thermal Modeling of Electron Devices.

As it has been stressed above, a research thrust has been enabled by the acquisition of the proposed equipment, aiming to the extension of the CMC modeling capabilities to the extraction of performance parameters (large DC and RF signal) that can be directly compared to experimental measurements. Within this approach, the device is modeled as a component of a larger system (a power amplifier), and its interaction with external components is the main object of the modeling effort. We refer to this project, which is currently funded by the ONR, as Thrust 1. The novelty of this approach is that *a)* it extends the nonlinear device simulation domain to include the reactive circuit surroundings of the intrinsic device and, most importantly, *b)* it extends the modeling capability well beyond the small-signal linear regime, making possible the realistic simulation of high-frequency and high-power devices. While retaining the well-established physical accuracy of full band particle-based Monte Carlo simulation method, the capability of our simulation tool will efficiently produce reliable and robust device parameter sets that will be efficiently and accurately included in coarse-grain circuit simulations for high speed and high power applications.

A second and third thrust of the research have been proposed and recently funded by DARPA, and will extend the CMC modeling capabilities on the opposite end of the simulation space: the microscopic interactions between charge carriers (electrons) and heat carriers (phonons). We offer here a description of the preliminary results of the two separate projects, aiming to the realization of self-consistent electro-thermal modeling of semiconductor devices (Thrusts 2 and 3). The first one is based on a representation of the process of heat generation and transport within the device though the solution of the energy-balance equation for phonons, while the second models the same process via a direct representation of phonons as particles. In both cases the thermal models will be self-consistently coupled with the CMC algorithm used to model charge transport.

We feel that the strict integration of the software produced by the three thrusts in the same software tool will constitute an unprecedented link between the parameter-free physical models of the CMC and the macroscopic parameters relevant for the design of high-power and high frequency devices. Such integration will be enabled by the acquired computational platform.

Energy Balance Equation for Phonons.

As mentioned above, in order to achieve efficient electro-thermal modeling capabilities in the existing CMC code, we plan to implement the models for heat generation and transport in two complementary ways. The first approach, here described as Thrust 2, is based on an efficient technique to reach the electro-thermal steady-state within the device layout. Such technique is based on the solution of the Energy Balance Equation (EBE) for phonons (flux-based) coupled self-consistently with the particle-based electron dynamics. Once the electro-thermal steady-state condition were reached, the temperature map supplied by the EBE solver will be replaced by a population of phonons and the particle-based phonon dynamics simulation engine will start in order to solve transients and non-equilibrium heat transport. The implementation of the particle-based models for heat transport will model the

heat flow as particles (phonons) and will be the object of the third and final thrust of the research.

We decided to obtain the flux-based analysis of heat generation and transport by solving the Energy Balance Equation for phonons rather than the Heat Transport Equation. This new approach would allow a more accurate temperature map by supplying a separate solution for each phonon mode (or group of modes), and is more suited to self-consistent coupling with the electron dynamics. The approach we developed starts from the phonon BTE (see equation 8 further below) to obtain the energy balance equation for each phonon mode μ:

$$\frac{\partial W_\mu}{\delta t} = -\nabla \cdot \mathbf{F}_W + \left.\frac{\partial W_\mu}{\delta t}\right|_{e-p} + \left.\frac{\partial W_\mu}{\delta t}\right|_{p-p} \qquad (3)$$

where $W_\mu(\mathbf{r},t) = \frac{1}{\Omega}\sum_{\mathbf{k}} E_\mu(\mathbf{k})\, f_\mu(\mathbf{r},\mathbf{k},t)$ is the ensemble energy in the volume Ω of the reciprocal space, $\mathbf{F}_\mu(\mathbf{r},t) = \frac{1}{\Omega}\sum_{\mathbf{k}} v(\mathbf{k}) E_\mu(\mathbf{k})\, f_\mu(\mathbf{r},\mathbf{k},t)$ is the energy flux, and the two partial derivatives of W_μ in the LHS of the equation represent the rate of change of the ensemble energy due to electron-phonon and phonon-phonon interaction, respectively.

At steady-state, the equation reads:

$$\nabla \cdot \left(k_\mu(T,\mathbf{r})\nabla T\right) = -\left(\left.\frac{\partial W_\mu}{\delta t}\right|_{e-p} + \left.\frac{\partial W_\mu}{\delta t}\right|_{p-p}\right) = -P_\mu(\mathbf{r}). \qquad (4)$$

In this case, $\mathbf{F}_\mu(\mathbf{r},t)$ has been approximated with the steady-state relation $\mathbf{F}_\mu(\mathbf{r}) = -k_\mu(T,\mathbf{r})\nabla T$, where k is the (scalar) thermal conductivity. The derived energy balance equation is indeed the classical heat transport equation; the novelty of the proposed approach is in the way the forcing function is computed within the Cellular Monte Carlo framework as well as the fact that one of such equation can be solved for each phonon mode.

The component of the forcing function $P_\mu(\mathbf{r}) = \left.\frac{\partial W_\mu}{\delta t}\right|_{e-p} + \left.\frac{\partial W_\mu}{\delta t}\right|_{p-p}$ due to electron-phonon scattering $\left.\frac{\partial W_\mu}{\delta t}\right|_{e-p}$ can be extracted in real time by recording the phonon emissions and absorptions for each mode μ after each electron-phonon scattering event, while the phonons decays and recombination events contributing to $\left.\frac{\partial W_\mu}{\delta t}\right|_{p-p}$ are approximated via relaxation time approximation. It should be noted that other approaches present in literature are based on a solution of the HTE and rely on one or more of three main assumptions: 1) the assumption that a specific energy decay path exist (electron → optical phonon → acoustic phonon), 2) the assumption that the acoustic phonon scattering are elastic, and 3) the assumption that the relaxation time approximation can be used also for determining $\left.\frac{\partial W_\mu}{\delta t}\right|_{e-p}$. All these approximations are not needed by the proposed treatment. In particular, our approach will produce the statistical relevance of each energy path (see assumption 1) above) as an output, rather than assuming a specific one. Finally, we note that even the relaxation time approximation for the phonon-phonon scattering can be abandoned once a particle-based approach will be completed for the phonon dynamics. At that moment, the $\left.\frac{\partial W_\mu}{\delta t}\right|_{p-p}$ term in the forcing function can be extracted and tabulated as a function of the temperature from a simple bulk simulation performed by using phonons as particles.

In principle, the nonlinear steady-state energy balance equation above can be solved via a 3D nonlinear finite element solver capable of handling the functional dependencies of $k_\mu(T, \mathbf{r})$. We decided not to pursue that approach for the following reasons: 1) such solver would have severe convergence problems and robustness issues due to the complexity of the geometry of the devices of interest, 2) the implementation of such solver is, per se, a formidable numerical problem, due to the necessity of generating an appropriate 2D or 3D grid for the devices, and 3) an extremely efficient finite-difference 2D and 3D multi-grid elliptical solver [32] is readily available within the existing code as discussed above. For these reasons, we further manipulated the energy balance equation in order to re-write it as an elliptical PDE.

The main issue with such manipulation is the dependency of k_μ from both the temperature and the position. We therefore assume that, with respect to the position, the thermal conductivity can be represented as a piece-wise function of the temperature, in other words, $k_{\mu,C}(T)$ is a function of the temperature but is not changing with the position within each cell *C* of the finite differences grid. We therefore express this restricted position dependency with the index *C* rather than via a full functional dependence on the position vector ***r***. Furthermore, we define an equivalent temperature $\theta_{\mu,C}(T)$ through the well-known Kirchhoff transformation [33]:

$$\theta_{\mu,C}(T) = T_0 + \frac{1}{k_{\mu,C}(T_0)} \int_{T_0}^{T} k_{\mu,C}(\tau) d\tau, \qquad (5)$$

Where T_0 is a reference temperature. This allows us to rewrite the energy balance equation as follows:

$$\nabla^2 \theta_{\mu,C} = -\frac{P_\mu(\mathbf{r})}{k_{\mu,C}(T_0)}, \qquad (6)$$

which is a linear Poisson equation for the transformed temperature $\theta_{\mu,C}(T)$. This equation can be easily and efficiently solved with the existing multi-grid 2D and 3D solver available in the existing code.

However, a crucial issue is related to the Kirchhoff transformation above, which needs further discussion. Indeed, while $k_{\mu,C}(T)$ is safely assumed not to change with position within a finite difference cell, the different materials present in a device layout are expected to have a different functional dependency on temperature. In other words, we need to assign different functions $k_{\mu,C}(T)$ in different cells. To achieve that, the temperature and its normal derivative need to be continuous across cell boundaries even when the thermal conductivity is not. This implies that, in order to have a unique solution of the linearized (elliptical) form of the energy balance equation, the Kirchhoff transformation must be invariant for the continuity conditions expressed above. The invariance means that if T and its normal derivative are continuous across cell boundaries, the equivalent temperature $\theta_{\mu,C}$ must have the same property. Unfortunately, while the Kirchhoff transformation is invariant for the continuity of the normal derivative of T, it is not invariant for the condition of continuity of T across a boundary [34]. This means that if T is continuous across a cell boundary, there is no guarantee that $\theta_{\mu,C}(T)$ is also continuous. This shouldn't be a surprise: indeed the application of the Kirchhoff transformation does not eliminate the non-linearity from the equation; it just moves it from inside the cell to the boundary of the cell [35]. In passing, one should note that if we would have chosen to use a non-linear finite element approach, we

would have moved the issue of non-linearity from the analytical domain to the numerical one because of the convergence issues of finite elements nonlinear solvers in the presence of complex boundaries. So the overall level of difficulty would not have changed.

In order to address this rather important limitation we note that for many different semiconductor materials, we can express their thermal conductivity $k_{\mu,C}(T)$ through the same functional form $f(T)$:

$$k_{\mu,C}(T) = \propto_C \ f(T), \qquad (7)$$

where $\propto_C$ is a constant that can vary in each cell. It is easy to verify that such condition translates in the needed invariance of the temperature continuity across boundaries. Fortunately, it turns out that this rather restrictive condition on the functional dependency can be successfully adopted for most semiconductors without significant loss of generality. As a final remark, we note that the condition on the functional form of $k_{\mu,C}(T)$ would be impossible to enforce at interfaces between semiconductors and metals, but metals are not directly simulated in the structures of interest.

Coupled Electron-Phonon Dynamics for Electro-Thermal Simulation. A second approach is described here, which is not based on the numerical solution of the EBE, but rather on the stochastic particle-based solution of the solution of the Boltzmann Transport Equation for phonons. In order to accurately simulate the heating, heat distribution and eventual thermal failure of semiconductor devices, it is critical to correctly understand and model at the microscopic scale the complex electro-thermal coupling effects, the resistive-thermal relationship in the devices, and their thermal breakdown property. Existing models cannot depict electro-thermal phenomena accurately because the phonon heating effect is not fully taken into account. However, it is believed that the phonon dynamics plays a main role in self-heating of devices, especially for sub-100nm and nano-structures where the ballistic-phonon scattering causes micro-over-heating as critical device dimensions become comparable to phonon mean free path. To fully account for the electro-thermal behavior, we propose to include full particle-based phonon dynamics into the electro-thermal device modeling framework.

In semiconductors, while most of the heat generation is due to the interaction between electrons and phonons, the heat conduction is mainly due to the motion of phonons. When device dimensions are larger than the phonon mean free path, heat transport occurs in a diffusive regime that is effectively described by the EBE. However, when the device size is comparable with the effective mean free path of phonons, the heat transport regime becomes ballistic, phonons undergo fewer scattering events, and the thermal transport occurs in non-equilibrium conditions. Being, for example, the phonon effective mean free path in silicon approximately 300nm [36], it is obvious that the heat flux in many current devices occurs within this non-equilibrium ballistic limit and that the EBE is inadequate to modeling the heat flux. To address this problem and in full analogy with the procedure used by the CMC approach in simulating charge transport, we will propose a thrust to model the phonon dynamics by the stochastic solution of the phonon Boltzmann equation:

$$\frac{\partial f_p(\mathbf{r},\mathbf{q},t)}{\partial t} = -V_p \cdot \nabla f_p(\mathbf{r},\mathbf{q},t) + \left[\frac{\partial f_p(\mathbf{r},\mathbf{q},t)}{\partial t}\right]_{coll} \qquad (8)$$

where f_p is the phonon distribution function expressed in terms of position $\mathbf{r}$, momentum $\mathbf{q}$ and time, V_p is the phonon group velocity, and the second term on the right side of the

equation is a collisional term accounting for the interactions of the phonons with other phonons, boundary conditions and electrons. The technique is inspired by the work of Mazdumer and Majumdar [37], and is based on the representation of the phase-space evolution of a significant portion of the phonon population and on its coupling with the charge carrier particles (electron and holes). In analogy with what is done with charge carriers, the dynamics of the phonons will be modeled as a sequence of ballistic free-flights interrupted by stochastic momentum changing collisions due to phonon-phonon, phonon-electron, phonon-impurity, and phonon-boundary scattering. The phonon scattering probabilities of three-phonon Normal and Umklapp scattering [38,39], as well as the phonon-impurity scattering, will be included in the fully discretized BZ of the CMC simulative framework by using perturbative analysis. In other words, we propose a stochastic self-consistent solution of both phonon and electron BTEs. A crucial aspect of this work will be the model used for the phonon-electron interaction. Within the proposed framework, the Boltzmann transport equation is expressed for electrons as follows:

$$\frac{\partial f_e(\mathbf{r},\mathbf{k},t)}{\partial t} = -V_e \cdot \nabla f_e(\mathbf{r},\mathbf{k},t) - \mathbf{F} \cdot \nabla_{\mathbf{k}} f_e(\mathbf{r},\mathbf{k},t) + \left[\frac{\partial f_e(\mathbf{r},\mathbf{k},t)}{\partial t}\right]_{coll} \qquad (9)$$

where the index *e* identifies quantities related to charge carriers. Note the drift term $\mathbf{F} \cdot \nabla_{\mathbf{k}} f_e(\mathbf{r},\mathbf{k},t)$, not present in the phonon BTE (equation 8), which accounts for the long-range electrostatic interaction mediated by the electrostatic force **F**. The collisional term for charge carriers $\left[\frac{\partial f_e(\mathbf{r},\mathbf{k},t)}{\partial t}\right]_{coll}$ accounts for the momentum-changing interactions of electrons and holes with their environment, including phonons. In particular, within the CMC framework, nonpolar phonon scattering is handled via perturbative approach in order to obtain the scattering rate from a region of band *n* centered in the point **k** to a region $\Omega_{\mathbf{k}'}$ in band *n'* centered around the point **k'** in momentum space:

$$P_{nn'}(\mathbf{k}_n, \Omega_{\mathbf{k}'}) = \frac{\pi}{\rho\omega_{\eta\mathbf{q}}} \left|\Delta_{\eta,n'}(\mathbf{q})\right|^2 \omega_{\eta\mathbf{q}} \left|I(n,n';\mathbf{k},\mathbf{k}')\right|^2 D_{n'}(E',\Omega_{\mathbf{k}'})(n_{\eta\mathbf{q}} + \frac{1}{2} \pm \frac{1}{2}) \qquad (10)$$

where ρ is the semiconductor density, $\mathbf{q}$ the phonon wave vector, $\Delta_{\eta,n'}(\mathbf{q})$ is the nonpolar matrix element, *I* is the overlap integral between Bloch states, $D_{n'}(E',\Omega_{\mathbf{k}'})$ is the electronic density of states in $\Omega_{\mathbf{k}'}$ at the final (after scattering) energy *E'* in band *n*', and $n_{\eta\mathbf{q}}$ is the phonon occupation number usually *computed at equilibrium* and *at the lattice temperature*. While this rate could be used "as is" for the calculations of the Joule term for the heat diffusion equation, a crucial modification is necessary in the non-equilibrium framework of the proposed particle representation for both phonons and charge carriers. The equilibrium phonon number will be replaced by the *local value* of the phonon population at the position where the scattering occurs. This value is supplied by the part of the algorithm that models the phonon dynamics as a solution of the phonon BTE. Also in this case, the scattering rates $P_{nn'}(\mathbf{k}_n, \Omega_{\mathbf{k}'})$ will be tabulated for the estimated maximum value of $n_{\eta\mathbf{q}}$ and a rejection technique will be implemented to handle storage issues. As a consequence of the scattering, a phonon creation/destruction mechanism will be implemented for the three-phonon scattering mechanism.

It should be stressed that the inclusion of the full-band representation for *both* electron and phonon dispersions is essential for a correct quantitative evaluation of the energy exchange between the particles in the system.

Some final considerations should be made concerning the performance of the CMC simulation tool. The method proposed here is based on the full self-consistent particle-based simulation of the charge and heat transport due to the combined dynamics of electrons and phonons. The time resolution of the algorithm is expressed in fractions of femtoseconds, while the timescale of the heat transfer from the electrons to the lattice is measured in fractions of picosecond [40] and the relaxation time of the phonon-driven thermal transport is approximately 80 picoseconds [41]. This means that a simulation should be carried for about 100 picoseconds in order to reach the electro-thermal steady state and successfully extract the device parameters to be used by the fast circuital simulators. By coupling the particle-based electron dynamics with the flux-based solution of EBE described in the previous section on Trust 2, we will be capable of efficiently reaching the electro-thermal steady-state condition. Successively, the simulation will be continued by performing a particle-based simulation of both electron and phonons. This multi-scale approach will allow studying the transient, non-linear electro-thermal properties of high-power and high-frequency devices.

5. Research Proposed to the DoD.

The design of the computational platform described in this document has been optimized for the particle-based CMC simulation of solid-state devices of strategic relevance. In general, the studies described in this document could not be performed with conventional state-of-the-art software/hardware, *i.e.* with software tools that are commercially available and can be executed on standard computer clusters. The acquisition of such equipment successfully allows the pursuit of three main research trusts in a timeframe that will span the entire useful life of the acquired equipment. While work for Thrust 1 has been started when funding was granted by the ONR, a brief summary of the two thrusts to be proposed for funding to the DoD is offered below:

Thrust 2) has been proposed for funding to DARPA in March 2014, funding has been granted in April 2015. *Self-consistent coupling of CMC and multigrid energy-balance solver for phonons*. Initial work for this thrust has been funded by DARPA and is described and motivated in the first part of section 4.

Thrust 3) has also been successfully proposed for funding to DARPA in March 2014 within the same proposal for Thrust 2). *Self-consistent coupling of electron and phonon particle-based dynamics in CMC simulation*. This project is discussed in the second part of section 4 and will complement the functionality achieved by Thrust 2). In other words, we plan to use the energy balance equation to thermalize a device and reach a self-consistent electro-thermal steady-state. The simulation will then continue by replacing the temperature map within the device with a phonon population in order to perform the study of high-speed transient behavior. The electro-thermal modeling capability granted by the software tools produced by Thrusts 2 and 3 will be determinant to increase the accuracy of the RF analysis of power amplifiers resulting from the Thrust 1) currently funded by the ONR. Initial work for this thrust has been funded by DARPA and is described and motivated in the second part of section 4.

It should be noted that the three trusts described above have a specific common characteristics: they all are aiming to the realization of realistic devices made of materials that have not been yet fully characterized experimentally. The main idea behind this

document is to complement, or even sometimes bypass, the low level microscopic experimental characterization of the materials by the integration of the somehow abstract predictions of *ab-initio* methods with state-of-the-art particle-based CAD tools.

6. New and Enhanced Research Capabilities

The acquired computational platform enhances the current research activities of the PI's group by dramatically increasing the number-crunching capability of his current computer systems. In particular, the computer-demanding optimization of material properties will be extended from binary compounds such as InN and AlN to ternary alloys. This allows, for example, the systematic study of the influence of the molar fraction on the transport characteristics in the alloyed channel of a HEMT. Another substantial enhancement of the current capability is related to the optimization of the layout of devices. Having many CPUs addressing a very large amount of RAM will allow the concurrent execution of several dozens of simulations to assist the crucial phase of optimization of novel device layouts in the presence of different concentrations of material defects.

Concerning the new research capabilities that have been acquired because of the new computational platform, its extremely large fast memory allows the modeling of transport occurring simultaneously on the different layers defining a heterojunction transistor. For the first time, we are able to study the effect of parasitic conduction paths defined on different materials, and their effect on the device figures of merit such as the transconductance, the cutoff frequency, and the maximum operational frequency. This acquired capability does not represent an incremental enhancement of our computational setup; it enables us to bring the modeling effort to an unprecedented level of accuracy and volume. This particularly applies to the studies of material-related reliability and longevity of semiconductor devices.

Furthermore, the ambitious project on microscopic heat management described above would not be possible with the current resources available to the group. Indeed, both the traditional architecture and the size of our previous computational platform did not allow the execution of the self-consistent fullband particle-based electro-thermal transport algorithms that we plan to realize by simulating concurrently both the electron and the phonon population within the semi-classical framework defined by the BTE. The acquired computational platform has been designed (see section 7 below) mainly with that application in mind.

Once more, we would like to stress the fact that the capabilities that we have achieved with the acquired platform are not realistically achievable with a standard, general-purpose computer clusters like the ones normally available in academia. Indeed, both the architecture and the network configuration of the acquired platform are carefully tuned for particle-based simulation as it is discussed in the next session.

Finally, the equipment acquisition allows the development of a unique set of computer codes that will be instrumental in the design and optimization of novel devices structures realized with nonconventional materials. Such devices are expected to be extremely relevant for the mission of the DoD because of the high-frequency, high-power operations that can be achieved with high reliability in maintenance-free space-based applications.

7. The Computational Platform

The CMC approach that we developed has two main modes of operation requiring two different configurations of the hardware. For this reason, the standard paradigm implemented in academic computer clusters is not adequate, and a rather unusual architecture is needed both in the nodes and in the communication network.

A first mode of operation is what we call “material setup phase”, and consists in the preliminary calculations of electronic structure, phonon spectra, and scattering table for the material(s) of interest. This phase is characterized by CPU intensive floating-point calculations and optimizations that determine many quantities as a function of their position within a discrete version of the Brillouin Zone of the momentum space. These calculations are loosely coupled, in the sense that each quantity can be concurrently computed in different points of the region of interest. Because of its characteristics, the material setup phase requires therefore many powerful CPUs linked in a high speed network in order to obtain all the material parameters in a realistic amount of time (48 hours max per material, using about 150 CPU cores).

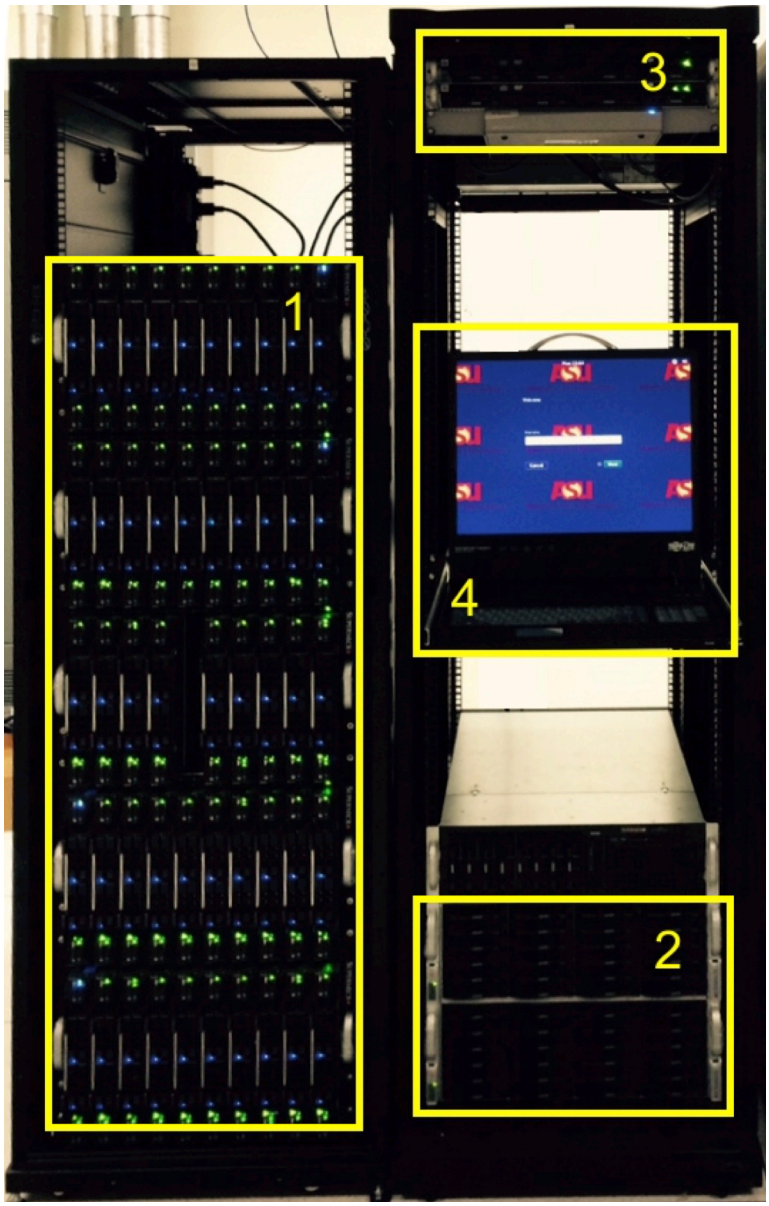

Figure 12. Main components of the acquired computational platform: 1) number-crunching nodes, 2) double-redundant distributed file server, 3) front-end nodes and fast network switches, 4) console.

The second, “simulation phase”, starts by loading the huge scattering tables computed in the previous phase and performs the simulation itself of a specific bias point of the device being studied. While each point is executed on an individual CPU, many CPUs can be used in parallel to obtain several bias points concurrently. Besides ensemble and time averages, most of the calculations in the simulation phase are executed with integer arithmetic and are based on large out-of-cache lookup tables. Therefore, this phase requires a large amount of extremely fast access memory and a dedicated network to upload large amount of data to a file server.

The different character of the two calculation phases requires an approach that is definitely unusual when designing the computer configuration.

Acquired Equipment

The computational platform acquired with funds granted by the AFOSR Grant FA9550-14-1-0083 is depicted in figure 12. Its four main components are highlighted and indicated with numbers, and are described as follows:

1. **Number-crunching nodes.** In its final configuration, the computing machinery includes 50 twin blades, each containing two independently addressable nodes. Each of the 100 nodes is equipped with two last-generation processors with a total of 16 cores operating at 2.6GHz. Each node is also equipped with 128 GB of DDR4 high-speed RAM and 1TB of local storage for the operating system and other local data. The 50 blades are grouped in 5 7U enclosures, each equipped with two high-speed (10Gbs) uplinks, connected to high-speed switches.
2. **File server.** A double-redundant distributed file server is made of two 4U enclosures; each equipped with 36 6TB hard drives connected in RAID5 configuration. Data on the disk bank of one enclosure are automatically replicated on the disks of the second enclosure so achieving double redundancy (RAID5 + mirroring). The distributed file system of choice is Gluster [42]. In its final configuration, the file server sustains about 0.18PB of double redundant storage over 0.43PB of the total "raw" capacity of the 72 disks. The Gluster file system is externally mountable by a variety of computer systems with the Samba [43] interoperability suite of programs. Stable Samba clients are available for Linux, Microsoft Windows, Apple OSX, Apple IOS, and Android computers.
3. **Front-ends and switches.** External access is granted by a front-end computer that establishes an encrypted Virtual Private Network (OpenVPN [44]) based on the Secure Socket Layer (SSL) paradigm. Each user is given a security certificate for each device used to connect to the computing nodes. Stable OpenVPN clients are available for Linux, Microsoft Windows, Apple OSX, Apple IOS, and Android computers. A backup front-end computer is present in case of failure of the main one. Two 16-port high-speed (10GBs) switches are used to connect two independent LANs -- one for disk sharing and one for inter-process message passing -- with the front-end computer that performs routing functions. Finally, a Keyboard-Video-Mouse (KVM) switch (previously acquired with other funding) allows the connection of each independent node of the system to a console for direct access.
4. **Console.** A simple LCD console (previously acquired with other funding) has been added to the system in order to access each node independently for direct maintenance.

The operating system installed on each node is CentOS 7.0 [45]. A local repository has been set in order to reduce the external bandwidth required by periodically updating more than 100 computing nodes.

The acquired equipment has been fully integrated in the existing computational infrastructure readily available to the PI. The new nodes have been integrated in the existing cluster used by the PI's group, and served by the same redundant distributed file server.

Given its size and structure, and relying on previous experience, we project the useful life of the acquired computational platform over a period of 5-7 years, longer than a typical industrial computer, and adequate for an academic one.

References

[1] "MEDICI 2D Semiconductor Device Simulation User's Manual", *Technology Modeling Association, Inc.*, CA, 1993.

[2] "Sentaurus reference Manual", 3rd ed., Synopsys, Montain View, CA,2007.

[3] "TSUPREM4 2D Semiconductor Process Simulation User's Manual", Technology Modeling Association, Inc., CA, 1993.

[4] R.W. Hockney and J.W. Eastwood, "Computer Simulation Using Particles,", Adam Hilger, Bristol, 1988.

[5] S. Selberher, "Analysis and Simulation of Semiconductor Devices," Springer-Verlag, Wien, 1984.

[6] N.W. Ashcroft and N.D. Mermin, "Solid State Physics", *Holt-Saumnders International*, Philadelphia, 1981.

[7] C. Jacoboni and P. Lugli, "The Monte Carlo Method for Semiconductor Device Simulation," Springer-Verlag, Wien, 1989.

[8] K. Kometer, G. Zandler, and P. Vogl, "Lattice–gas cellular–automaton method for semiclassical transport in semiconductors," *Physical Review B*, vol. 46, no. 3, pp. 1382–1394, July1992.

[9] C. Kittel, " Introduction to Solid State Physics,"Wiley, New York, 1971.

[10] M. Saraniti, S. J. Wigger, and S. M. Goodnick, "Full-Band Cellular Automata for Modeling Transport in Sub-Micrometer Devices," *Proc. 2nd Int'l Conf. on Modeling and Simulation of Microsystems*, pp.380-383, 1999.

[11] M.V. Fischetti and S.E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Physical Review B, Vol. 38, No. 14*, pp. 9721-9745, November 1988.

[12] M. Saraniti and S. M. Goodnick, "Hybrid Fullband Cellular Automaton/Monte Carlo Approach for Fast Simulation of Charge Transport in Semiconductors," *IEEE Trans. Elec. Dev. Vol. 47, No.10*, pp. 1909-1906, 2000.

[13] E. Merzbacher, "Quantum Mechanics," 2nd ed. , Wiley International, New York, 1961.

[14] M. Saraniti, J. Tang, S. Goodnick, and S. Wigger, " Numerical challenges in particle-based approaches for the simulation of semiconductor devices," Mathematics and Computers in Simulations 65, 501, 2003.

[15] W. Hackbush, "Multigrid Methods and Applications," Springer-Verlag, Berlin, 1985.

[16] M. Saraniti, A. Rein, G. Zandler, P. Vogl, and P. Lugli, " An efficient multigrid Poisson solver for device simulations," IEEE Trans. On CAD, vol. 15, no. 2, pag. 141, 1996.

[17] S.J. Wigger, "Three Dimensional Multigrid Poisson Solver for Use in Semiconductor Device Modeling," *M.S. Thesis*, EE Department, Arizona State University, Tempe, AZ, 1998.

[18] P. Wesseling, "Theoretical and Practical Aspects of a Multigrid Method," SIAM J. Sci. Stat. Comp., vol. 3, no. 4, pag. 387, 1982.

[19] R.W. Hockney, A fast direct solution of Poisson's equation using Fourier analysis. *J. Assoc. Comput. Mach.* **8**, pp. 95–113, 1965.

[20] P. Wesseling, "Theoretical and Practical Aspects of a Multigrid Method," SIAM J. Sci. Stat. Comp., vol. 3, no. 4, pag. 387, 1982.

[21] P.W. Hemker, "The incomplete LU-decomposition as a relaxation method in multigrid algorithms," pp. 306-311 in Boundary and interior layers - Computational and asymptotic methods, ed. J.J.H. Miller, Boole Press, 1980.

[22] P.W. Hemker, "On the comparison of line-Gauss-Seidel and ILU-relaxation in multigrid algorithms," pp. 269-277 in Computational and asymptotic methods for boundary and interior layers, ed. J.J.H. Miller, Boole Press, 1982.
[23] P.W. Hemker, R. Kettler, P. Wesseling, and P.M. de Zeeuw, "Multigrid methods: development of fast solvers," Appl. Math. Comp. 13 pp. 311-326 1983.
[24] R. Quay, "Gallium Nitride Electronics," no. 96 of Springer Series in material Science, Springer, Berlin, 2008.
[25] C. Look, J. R. Sizelove, "Dislocation Scattering in GaN," Physical Review Letters 82, n. 6, pp. 1237-1240 (1999).
[26] N. G. Weimann, L. F. Eastman, D. Doppalapudi, H. M. Ng, T. D. Moustakas," Scattering of electrons at threading dislocations in GaN", Journal of Applied Physiscs 83, n. 7, pp. 3656-3659 (1998).
[27] C. Lee, P. Saunier, J. Yang, M. Asif Khan, "AlGaN-GaN HEMTs on SiC with CW power performance of >4 W/mm and 23% PAE at 35 GHz", IEEE Electron Device Letters 24, pp. 613-615, October 2003.
[28] O. Ambacher, J. Smart, J. R. Shealy, N. G. Weimann, K. Chu, M. Murphy, W. J. Schaff, L. F. Eastman, R. Dimitrov, L. Wittmer, M. Stutzmann, W. Rieger, and J. Hilsenbeck, "Two dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN heterostructures", *J. App. Phys.,* vol. 85, pp. 3222-3233, Mar. 1999.
[29] G. Koley and M. G. Spencer, "Surface potential measurements on GaN and AlGaN/GaN heterostructures by scanning Kelvin probe microscopy", *J. Appl. Phys.,* vol. 90, pp. 337-344, July 2001.
[30] A. Raffo, S. D. Falco, V. Vadala', and G. Vannini, "Characterization of GaN HEMT low-frequency dispersion through a multiharmonic measurement system," IEEE Transactions on Microwave Theory and Techniques, vol. 58, no. 9, pp. 2490 – 2496, September 2010.
[31] H. I. Fujishiro, S. Narita, and Y. Tomita, "Large signal analysis of AlGaN/GaN-HEMT amplifier by coupled physical device-circuit simulation," Physica Status Solidi (a), vol. 203, no. 7, pp. 1866 – 1871, 2006.
[32] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, (1985).
[33] H.S. Carlslaw and J.C. Jaeger, *Conduction of Heat in Solids*, Oxford Univ. Press, (1959).
[34] F. Bonani and G. Ghione, Solid-State El., **38**, 7, pp.1409-1412, (1995).
[35] F. Bonani, *private communication*, (2011).
[36] Y.S. Ju and K.E. Goodson, "Phonon Scattering in Silicon Films with Thickness of Order 100 nm," *Applied Physics Letters, Vol. 74, No. 20*, pp.3005-3007, 1999.
[37] S. Mazumder and A. Majumdar, "Monte Carlo Study of Phonon Transport in Solid Thin Films Including Dispersion and Polarization," *Journal of Heat Transfer, Vol. 123,* pp.749-759, August 2001.
[38] M.G. Holland, "Analysis of Lattice Thermal Conductivity," *Physical Review, Vol. 132, No. 6*, pp. 2461-2471, 1963.
[39] M.G. Holland, "Phonon Scattering in Semiconductors from Thermal Conductivity Studies," *Physical Review, Vol. 134, No. 2A*, pp. A471-A480, 1964.
[40] A. Majumdar, *Microscale Energy Transport*, C.-L. Tien, et. al., eds., Taylor & Francis, New York, NY, 1998.
[41] Y.S. Yu, "Microscale Heat Conduction in Integrated Circuits and Their Constituent Thin Films," *Ph.D. thesis*, Stanford University, Stanford, CA, 1999.

[42] http://www.gluster.org
[43] https://www.samba.org
[44] https://openvpn.net
[45] https://www.centos.org

1.

1. Report Type

Final Report

Primary Contact E-mail
Contact email if there is a problem with the report.

marco.saraniti@asu.edu

Primary Contact Phone Number
Contact phone number if there is a problem with the report

480-965-2650

Organization / Institution name

Arizona State University

Grant/Contract Title
The full title of the funded effort.

A Dedicated Computational Platform for Cellular Monte Carlo T-CAD Software Tools

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-14-1-0083

Principal Investigator Name
The full name of the principal investigator on the grant or contract.

Dr. Marco Saraniti

Program Manager
The AFOSR Program Manager currently assigned to the award

Dr. Kenneth Goretta

Reporting Period Start Date

04/15/2014

Reporting Period End Date

04/14/2015

Abstract

We report here on the acquisition of a specific computational platform with an optimized architecture for the Cellular Monte Carlo particle-based T-CAD simulation tools developed by our group. Such code is used for the modeling and design of electron devices realized with nontraditional semiconductor materials, where the experimental determination of material parameters for simulation purposes is arduous or incomplete.

A well-defined application space has been defined, in which the particle-based paradigm plays a crucial role for the design, realization, and optimization of electron devices of strategic relevance for present and future military applications. To be fully developed in a realistic amount of time, such application space requires a computational platform specifically designed for the algorithmic characteristic of particle-based simulation code. The application space ranges from the nano-scale first principles used to develop the code, to the extraction of macroscopic performance metrics of high-frequency, high-power electron devices in a strictly integrated simulation framework.

The acquired equipment has been successfully deployed and is making possible the extraction of device parameters from simulation data performed at molecular resolution. For the first time, the parameters extracted with our tools will supply quantitative predictions of direct RF measurements as well as the

electro-thermal device characteristics; the equipment will therefore allow the device design even in absence of reliable experiments.

In its final configuration, the computational platform is composed of 100 number-crunching nodes, each equipped with 16 computing cores and 128GB of fast storage (last generation DDR4 Random Access Memory). The 1,600 cores are connected to two high-speed local area networks (LANs) with 10Gb/sec bandwidth. The first LAN is dedicated to inter-process communication based on the Message Passing Interface (MPI) paradigm, implemented by the OpenMPI library. The second LAN connects each node to a double-redundant file server with 170TB storage capacity. Two automated backup appliances are located in different buildings and ensure daily snapshots of the most relevant data contained in the file server.

Access to the computational platform is granted by an encrypted connection base on the Secure Socket Layer (SSL) protocol, and implemented in the OpenVPN Virtual Personal Network communication software. A security certificate for each device used to connect to the computational platform is supplied to each user. Connections are allowed from MS Windows computers, OSX Apple computers, and Linux-based computers. Mobile devices running the iOS operating system from Apple are allowed as well, such as iPad and iPhone.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

FA9550-14-1-0083_Saraniti_SF298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

FA9550-14-1-0083_Saraniti_final_report.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Changes in research objectives (if any):

Change in AFOSR Program Manager, if any:

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Jul 08, 2015 14:35:54 Success: Email Sent to: marco.saraniti@asu.edu